Cloud Computing for Earth Sciences: Deployment of GEOSS Clearinghouse on Amazon's EC2

Qunying Huang[1], Chaowei Yang[1], Doug Nebert [2]

Kai Liu[1], Huayi Wu[1]

[1]Joint Center of Intelligent Computing

George Mason University

[2]Federal Geographic Data Committee (FGDC)

ESIP 2011 Winter Meeting

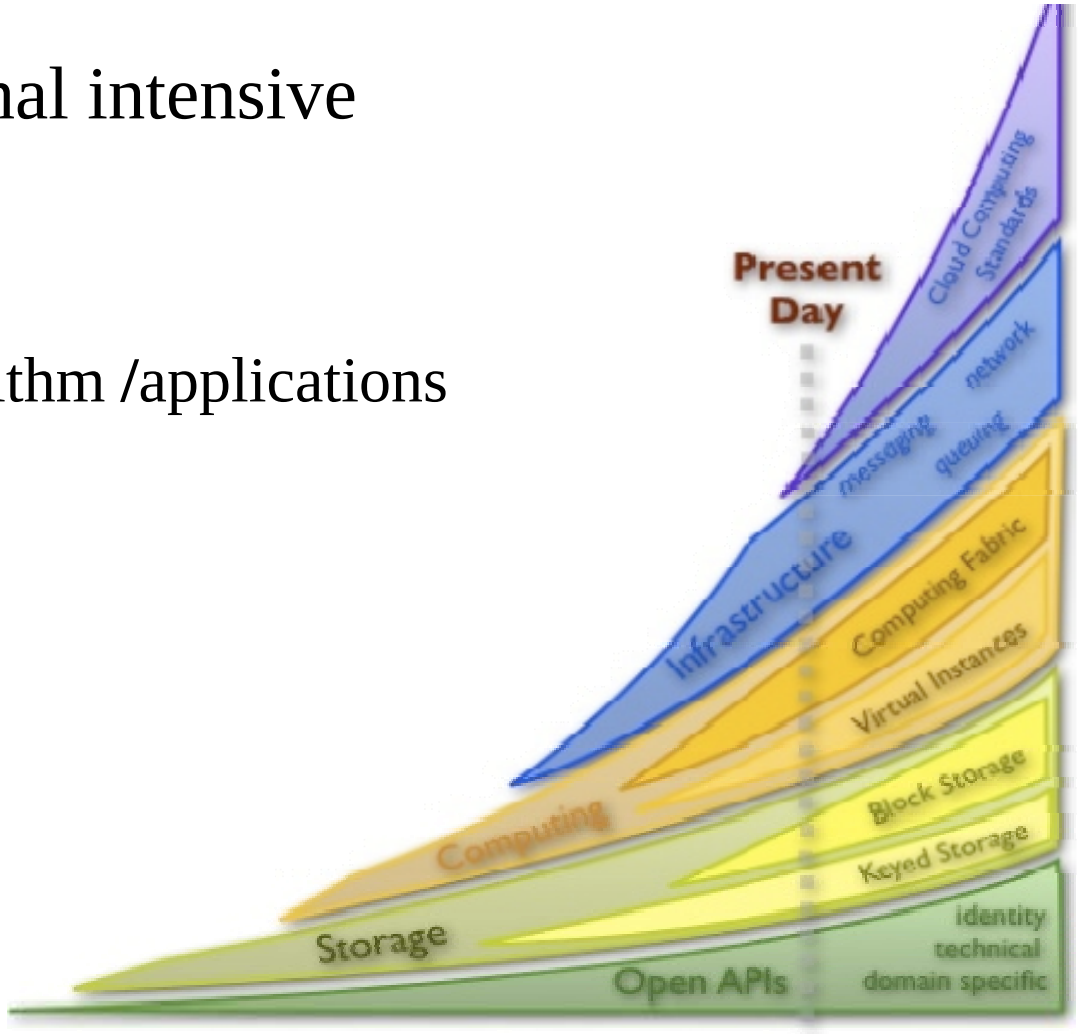
Jan 4th , 2011

Outline

- Introduction
- Related Work
- Deploying GEOSS Clearinghouse onto Amazon EC2
- Conclusion

Introduction

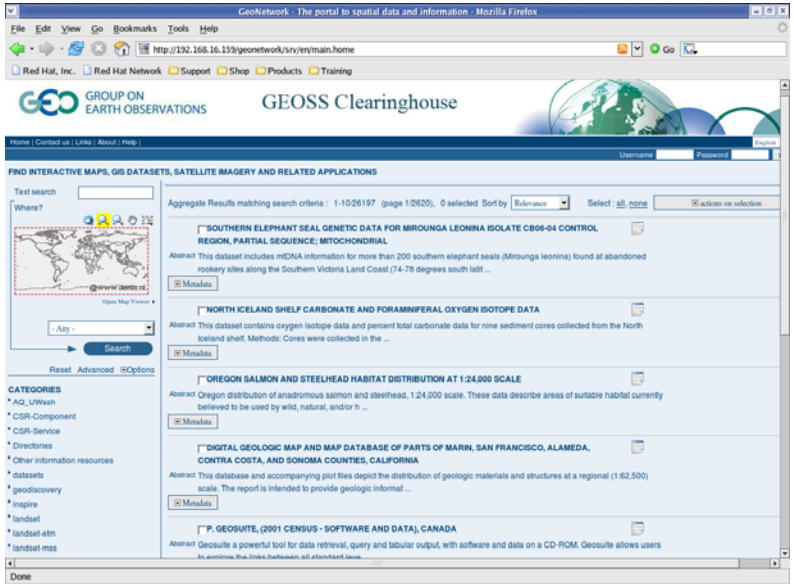
- Data and computational intensive
  - Scientific problems
  - Spatial analysis /algorithm /applications
- Computing paradigm
  - Cluster computing
  - Grid computing
  - Cloud Computing



The growth of cloud computing
From http://www.zdnet.com/blog/hichecliffe

Project Objectives

- GeoCloud
  - Ten geospatial application projects in the Cloud environment
  - Common operating system and software suites
  - Deployment and management strategies
  - Usage and costing of Cloud services
  - Security
- GEOSS Clearinghouse
  - Metadata catalogues search facility for the Intergovernmental Group on Earth Observation (GEO).



Cloud Computing

- Definition
  - “A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction “(NIST, 2010)
- Defining characteristics
  - On-demand self-service
  - Multi-tenancy
  - Measured Services
  - Device and Location independent resource pooling
  - Rapid elasticity

Cloud Computing Services

Software as a Service (SaaS)

- Almost any IT services
- Users: End-user

Platform as a Service (PaaS)

- Platform for developing and delivering applications, abstracted from infrastructures
- Users: Developer

Infrastructure as a Service (IaaS)

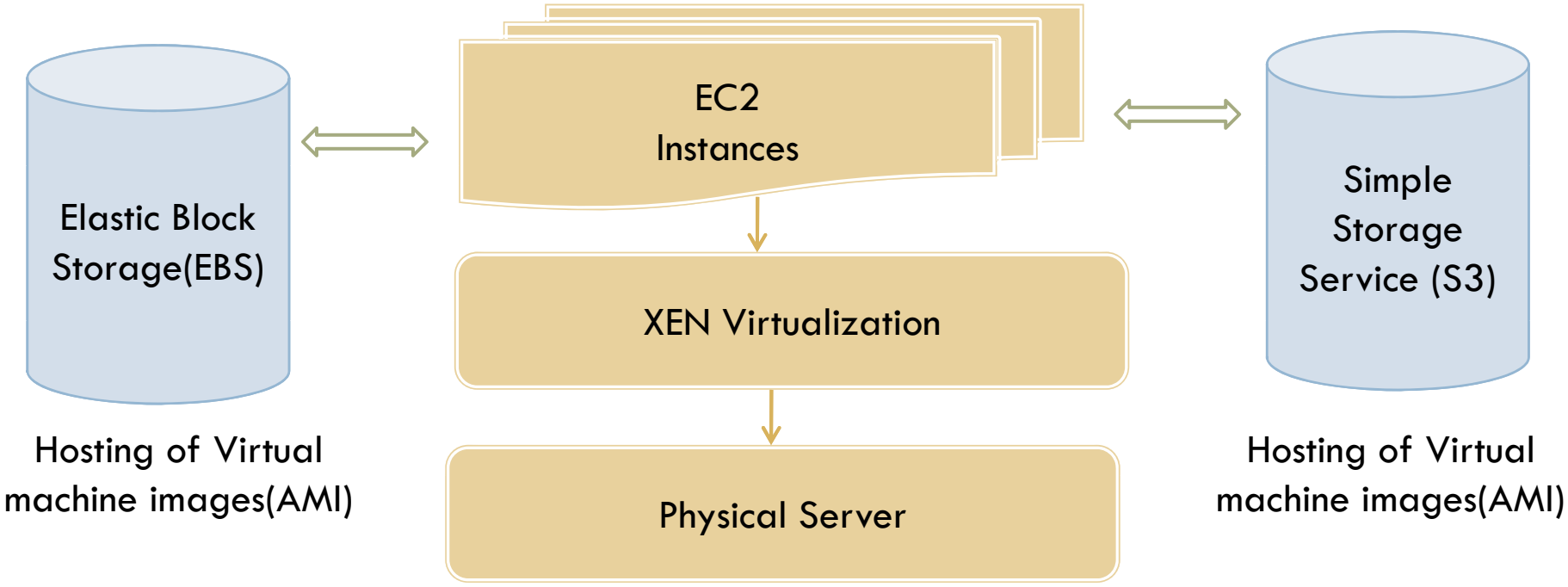
- On-demand sharing physical infrastructures
- Users: System Administrator

Amazon Cloud Services

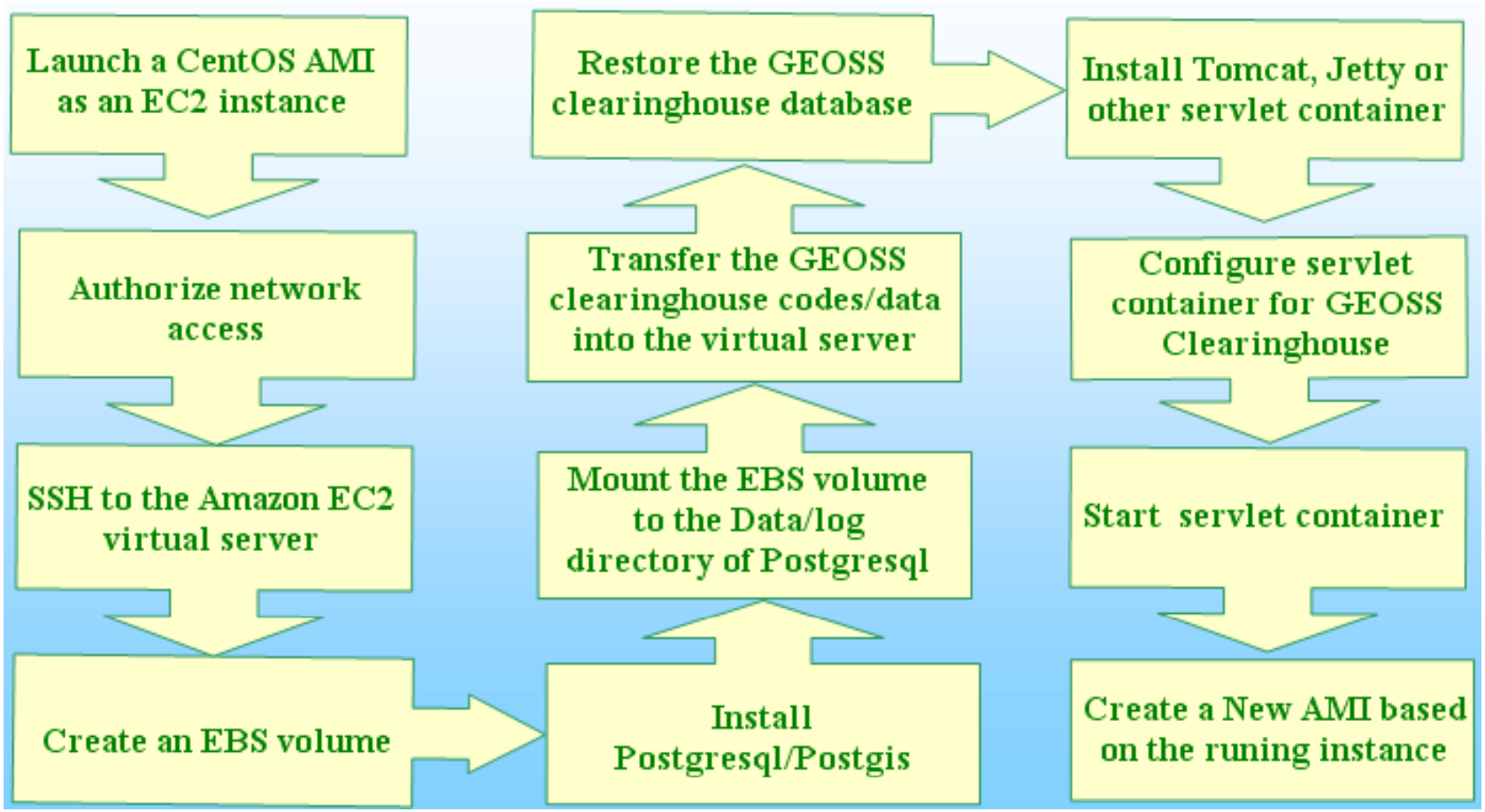
- Elastic Compute Cloud – EC2 (IaaS)
- Simple Storage Service – S3 (IaaS)
- Elastic Block Storage – EBS (IaaS)
- SimpleDB (SDB) (PaaS)
- Simple Queue Service – SQS (PaaS)
- Consistent AWS Web Services API (SaaS)

Amazon EC2

- A “Web service that provides resizable compute capacity in the cloud”
- Amazon Machine Image (AMI): a bootable VM image, which can be launched as a EC2 instance



Deployment of GEOSS Clearinghouse on Amazon EC2



Deployment of GEOSS Clearinghouse on Amazon EC2 -*cont*

- Scalability
  - Load balancer
- Reliability
  - Network
  - Disaster Recovery
- Reducing duplicated efforts
  - Infrastructure
  - Development

Amazon EC2 Standard Linux Instance Types

Type	CPU	Memory	Storage (unformatted)	Platform	I/O	AWS Name	Cost/ hour
Small	1 EC2-CU (1 virtual core with 1 EC2 Compute Unit)	1.7 GB (917MB swap)	170GB instance storage (160GB plus 10GB root partition, 1 spindle)	32-bit	Moderate	m1.small	$0.085 $747 a year or $490.30 a year Reserved
Large	4 EC2-CU (2 virtual cores with 2 EC2 Compute Units each)	7.5 GB (No swap)	910GB instance storage (2 x 450 GB plus 10GB root partition, 3 spindles).	64-bit	High	m1.large	$0.34 $2978 a year or $1961 a year Reserved
Extra Large	8 EC2-CU (4 virtual cores with 2 EC2 Compute Units each)	15 GB (No swap)	1810GB instance storage (4 x 450GB plus 10GB root partition, 5 spindles).	64-bit	High	m1.xlarge	$0.68 $5957 a year or $3922 a year Reserved

Amazon EC2 High-Memory Linux Instance Types

Type	CPU	Memory	Storage	Platform	I/O	AWS Name	Cost
High-Memory Extra Large	6.5 ECU (2 virtual cores with 3.25 EC2 Compute Units each)	17.1 GB	420 GB	64-bit	High	m2.xlarge	$0.50 per hour
High-Memory Double Extra Large	13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each)	34.2 GB	850 GB	64-bit	High	m2.2xlarge	$1.0 per hour
High-Memory Quadruple Extra Large	26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each)	68.4 GB	1690 GB	64-bit	High	m2.4xlarge	$2.0 per hour

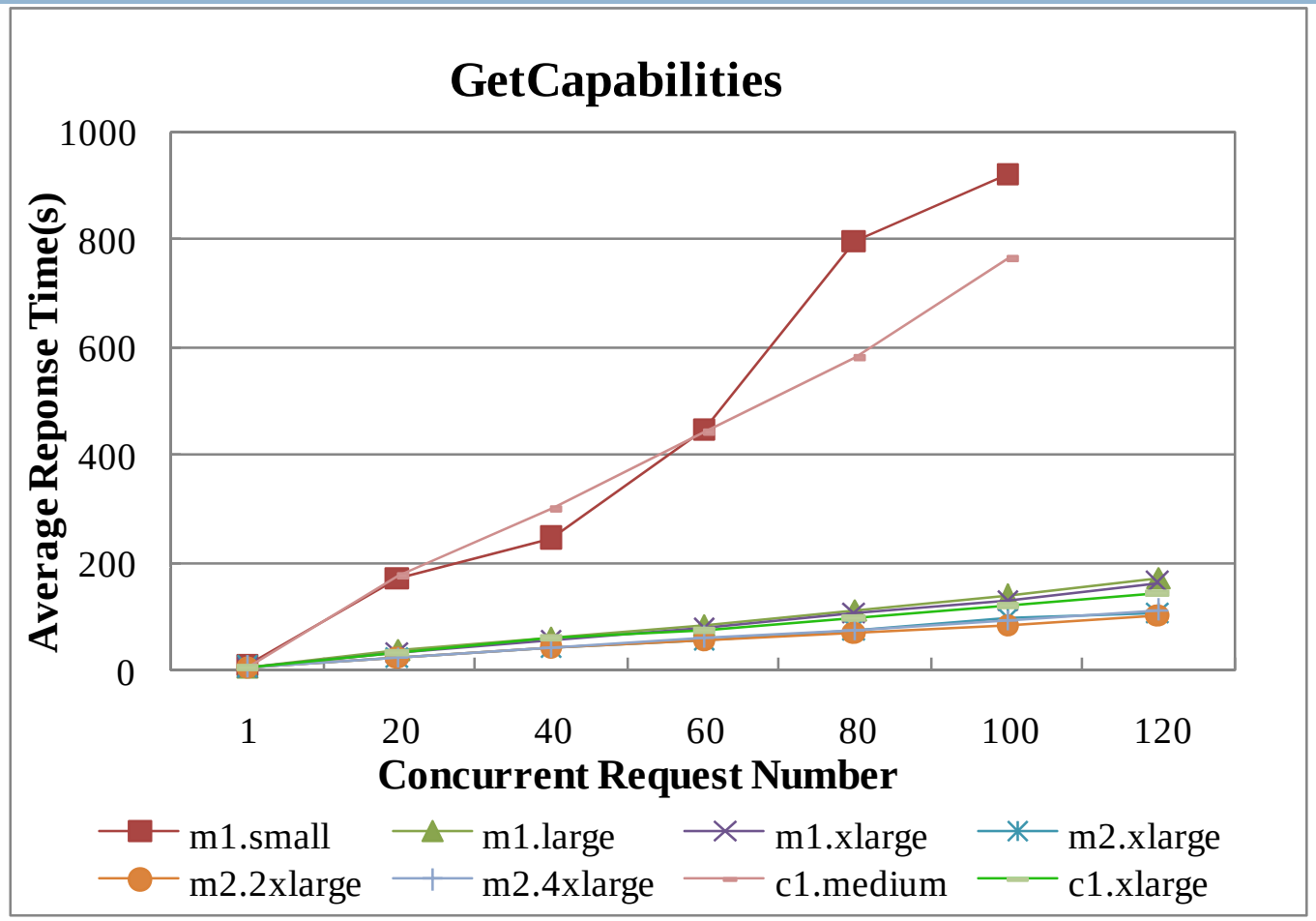
Amazon EC2 High-CPU Linux Instance Types

Type	CPU	Memory	Storage	Platform	I/O	AWS Name	Cost
High-CPU Medium	5 ECU (2 virtual cores with 2.5 EC2 Compute Units each)	1.7 GB	370 GB	32-bit	Medium	c1.medium	$0.17per hour
High-CPU Extra Large	20 Compute Units (8virtual cores with 2.5 EC2 Compute Units each)	7.5 GB	1810 GB	64-bit	High	c1.xlarge	$0.68 per hour

Amazon EC2 New Instance Categories

- Micro On-Demand Instances
  - Micro $0.02 per hour
- Cluster Compute Instances
  - 10 Gigabit Ethernet
  - Quadruple Extra Large $1.60 per hour
- Cluster GPU Instances
  - Quadruple Extra Large $2.10 per hour

Amazon EC2 Instance Performance Test



- GetCapabilies request from different number of concurrent requests

Conclusion

- Cloud computing
- We are at a prescient time
  - Technologies
  - Cloud Architecture
  - Platform independent languages
  - Open data standards
- Spatial Cloud Computing
  - Geospatial Middleware

Reference

- Huang Q., Yang C., Nebert D., Liu K., Wu H., 2010. Cloud computing for geosciences: deployment of GEOSS clearinghouse on Amazon's EC2, *Proceedings of ACM GIS HPDGIS workshop*, 35-38.
- Armbrust, M., Fox, A. and Griffith, R., et al. 2009. Above the Clouds: A Berkeley View of Cloud Computing. *Tech. Rep.* UCB/EECS-2009-28, EECS Department, University of California, Berkeley, CA, 2009. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html. (Accessed March, 2010).
- NASA Cloud Computing Platform. http://www.nasa.gov/offices/ocio/ittalk/06-2010_cloud_computing.html. (Accessed March, 2010).
- Huang Q. and Yang C. 2010. Optimizing Grid Computing Configuration and Scheduling for Geospatial Analysis-- An Example with Interpolating DEM. Computers & Geosciences, doi:10.1016/j.cageo.2010.05.015.
- Sun White Paper, 2009. Introduction to Cloud Computing Architecture, 1 edition, June, 2009.
- Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., and Warfield, A. 2003. Xen and the art of virtualization. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (Bolton Landing, NY, USA, October 19 - 22, 2003). SOSP '03. ACM, New York, NY, 164-177. DOI= http://doi.acm.org/10.1145/945445.945462
- Amazon Web Service. Available at: http://aws.amazon.com. (Access June, 2010).
- Lucene. Available at: (http://lucene.apache.org/java/docs/index.html) (Access Sep, 2010).
- Pavlo, A., Paulson, E., Rasin, A., Abadi, D. J., DeWitt, D. J., Madden, S., and Stonebraker, M. 2009. A comparison of approaches to large-scale data analysis. In *Proceedings of the 35th SIGMOD international Conference on Management of Data* (Providence, Rhode Island, USA, June 29 - July 02, 2009). C. Binnig and B. Dageville, Eds. SIGMOD '09. ACM, New York, NY, 165-178. DOI= http://doi.acm.org/10.1145/1559845.1559865
- Yang C., Raskin R., Goodchild M.F., Gahegan M., 2010a, Geospatial Cyberinfrastructure: Past, Present and Future, *Computers, Environment, and Urban Systems*, 34(4):264-277.
- Yang, C., Wu. H., Huang, Q., Li, Z. and Li, J., 2010b. Spatial Computing for Supporting Physical Sciences. *PNAS*. (in press).
- Armbrust, M. et al. 2010. A view of cloud computing. *Communications of the ACM*, 53(4): 50-58.

Thank You!

Qunying Huang

qhuang1@gmu.edu

Chaowei Yang

cyang3@gmu.edu

Doug Nebert

ddnebert@fgdc.gov

CISC

http://cisc.gmu.edu

Pointers

Portal
http://aws.amazon.com

Blog
http://aws.typepad.com

EC2
http://aws.amazon.com/ec2

S3
http://aws.amazon.com/s3

Resource Center
http://aws.amazon.com/resources

Forums
http://aws.amazon.com/forums